Data & Visual Analytics

Duen Horng (Polo) Chau
Georgia Tech

CSE6242 / CX4242

Jan 7, 2014

Who Am I?

www.cc.gatech.edu/~dchau/

Georgia Tech | College of Computing

CV (PDF) | Bio | Publications | Design | Twitter | Blog | Address

POLO CHAU

Legal name: Duen Horng Chau

Assistant Professor, School of Computational Science & Engineering
Adjunct Assistant Professor, School of Interactive Computing
College of Computing
Georgia Tech

polo@gatech.edu www.cc.gatech.edu/~dchau
Office: Klaus 1324 404-385-7682
Google Scholar (h-index: 17) YouTube videos

POSITIONS

Aug 2012 - Assistant Professor
School of Computational Science & Engineering, Georgia Tech

Dec 2012 - Adjunct Assistant Professor
School of Interactive Computing, Georgia Tech

EDUCATION

Aug 2012 **Ph.D. Machine Learning** Carnegie Mellon University
Thesis: Data Mining Meets HCI: Making Sense of Large Graphs
Award: Carnegie Mellon SCS Dissertation Award, Honorable Mention
Committee: Christos Faloutsos, Jason Hong, Niki Kittur, Jiawei Han (UIUC)

I'm teaching CSE 6242 / CX 4242 "Data and Visual Analytics" in Spring 2014.

Submit to our WWW 2014 Big Graph Mining workshop!

Attend our Big Graph Mining tutorial at WSDM'14, New York.

Our KDD'13 IDEA workshop was a great success!

Nov Invited talk at Rutgers University (11/5), Bell Labs (11/4)

Oct Attend DARPA ADAMS PI meeting, WIN'13, BigData'13. Invited speaker at Vis'13 workshop

Sept Two ICDM'13 demos accepted. Two WIN'13 papers accepted. Minimalist graph mining

Course Staff

Instructor
Duen Horng (Polo) Chau
Assistant Professor, CSE
Office hour: Thu 3-4pm, Klaus 1324

TA
Robert Pienta, PhD student, CSE

TA
Long Tran, PhD student, CS

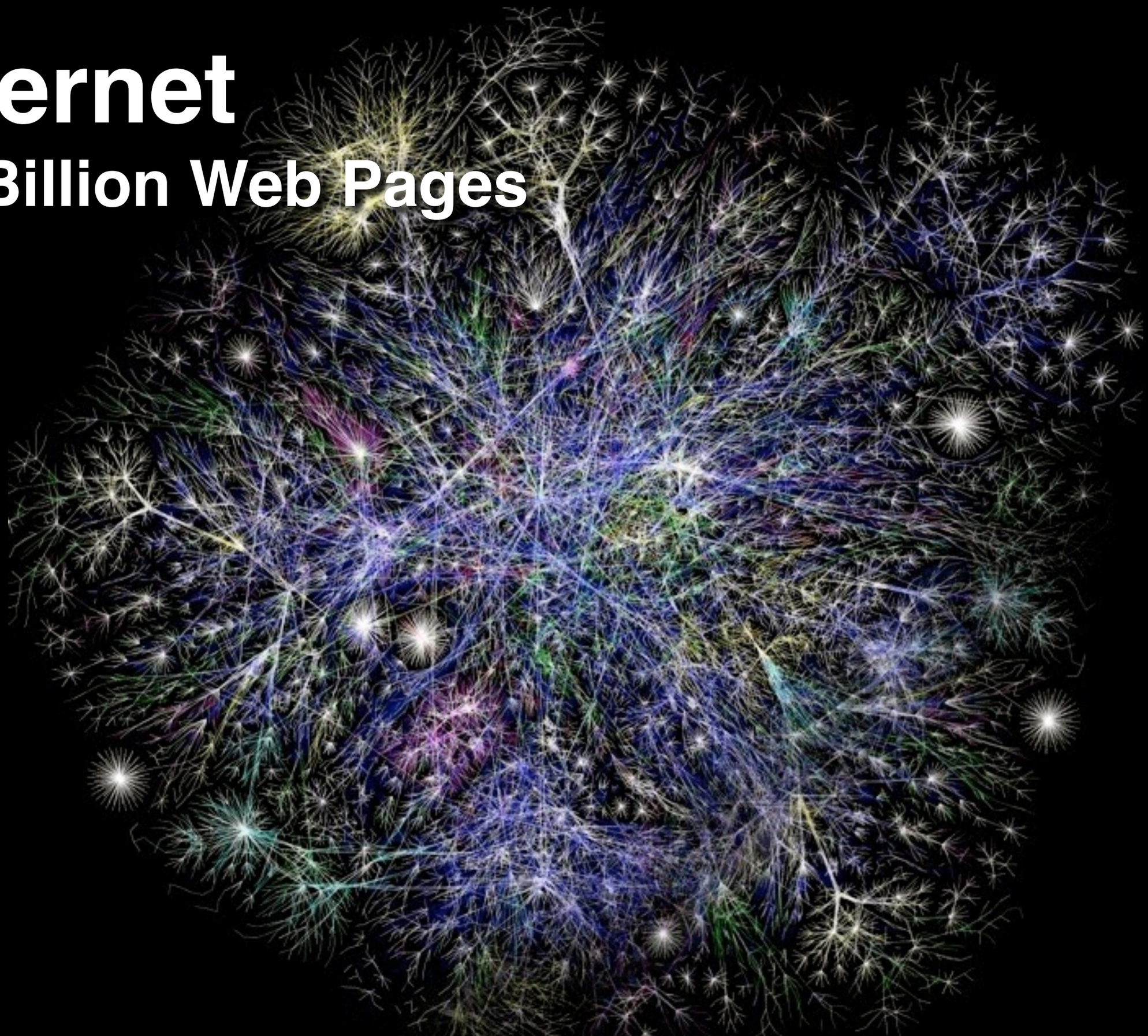
I Work with Large Graphs

I Work with Large Graphs

= Large Network Data

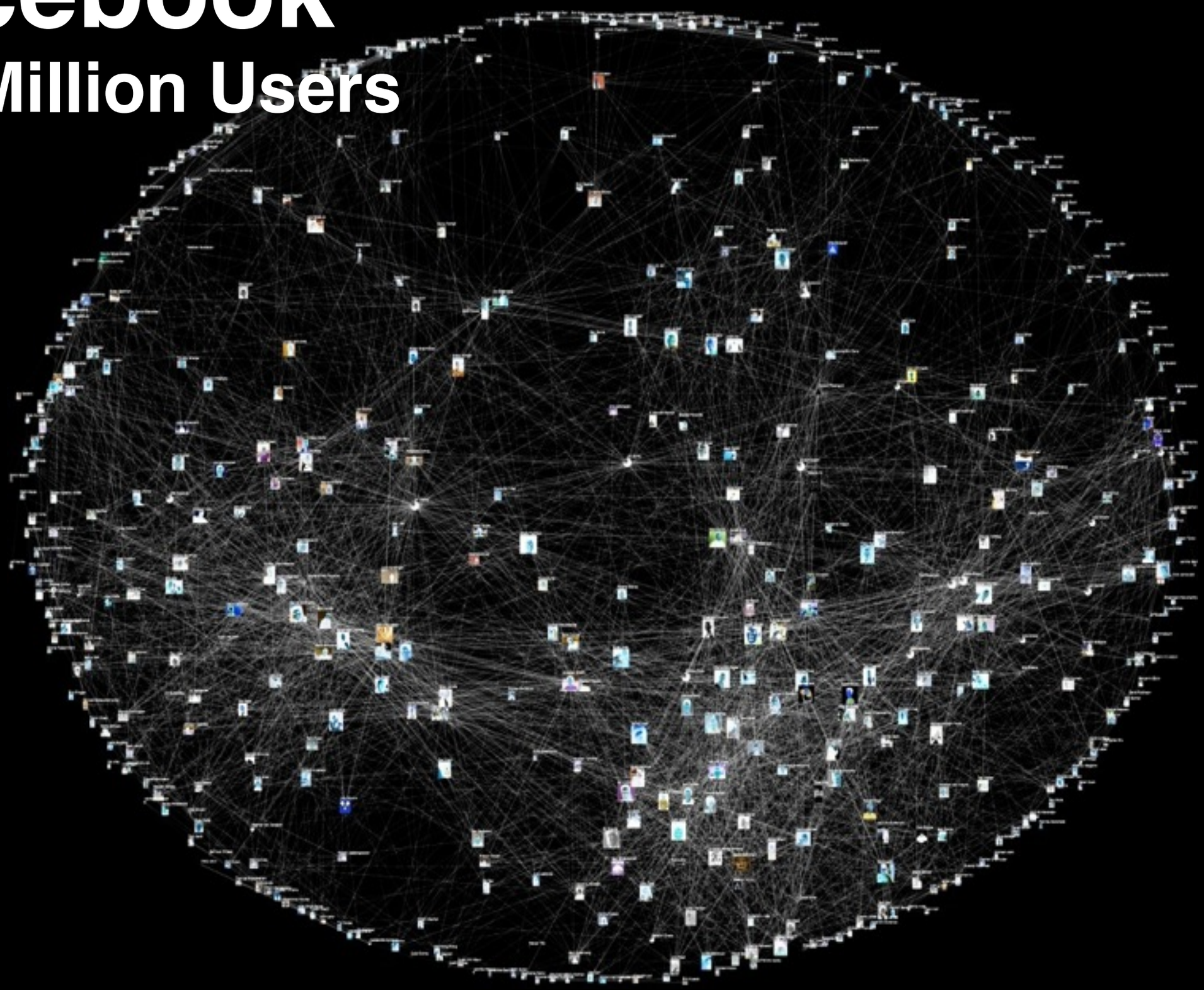
Internet

50 Billion Web Pages



www.worldwidewebsize.com www.opte.org

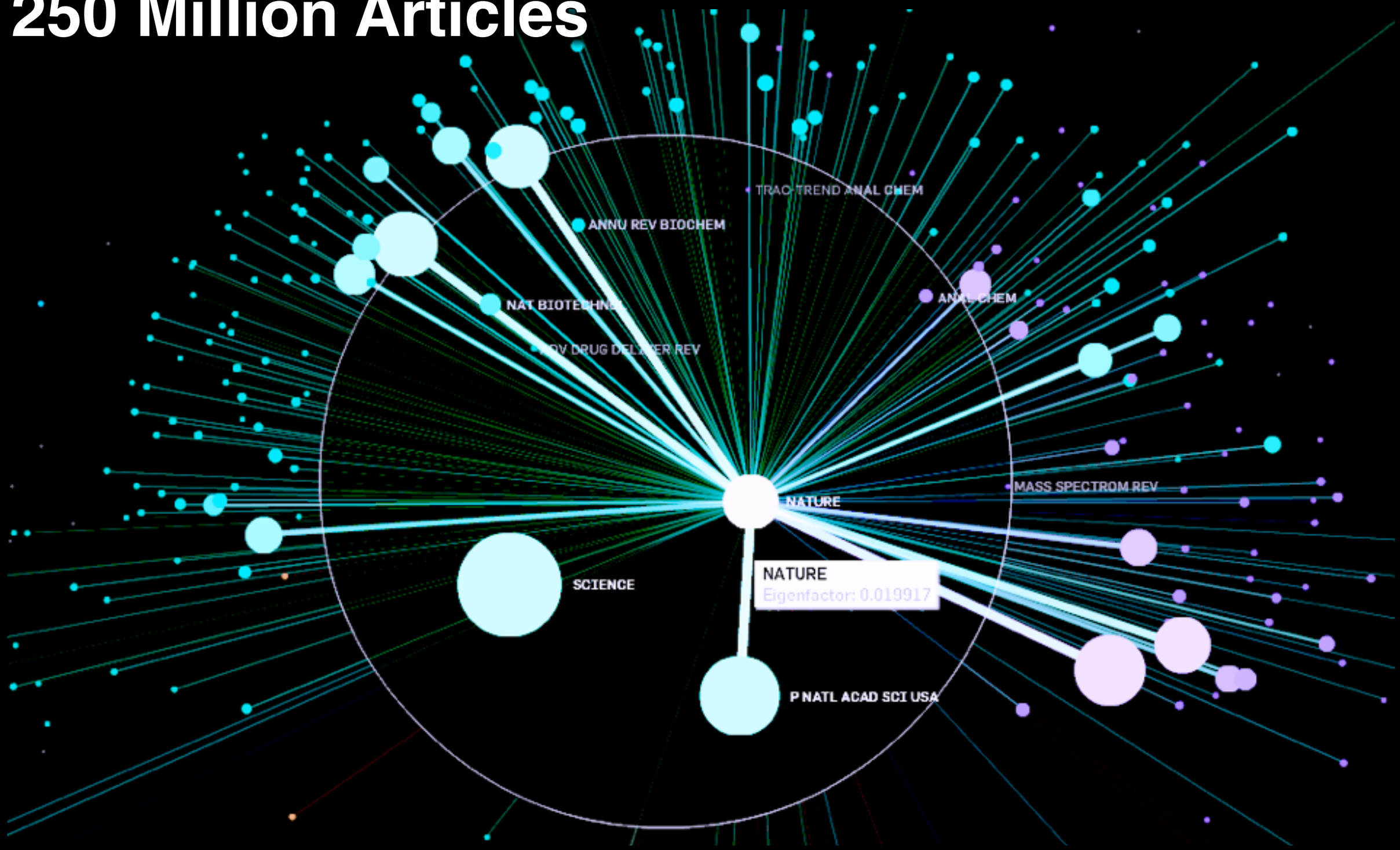
Facebook
800 Million Users



Modified from Marc_Smith, flickr

Citation Network

250 Million Articles



www.scirus.com/press/html/feb_2006.html#2 Modified from well-formed.eigenfactor.org

Many More

twitter

Who-follows-whom (500 million users)

amazon

Who-buys-what (120 million users)

at&t **cellphone network**

Who-calls-whom (100 million users)

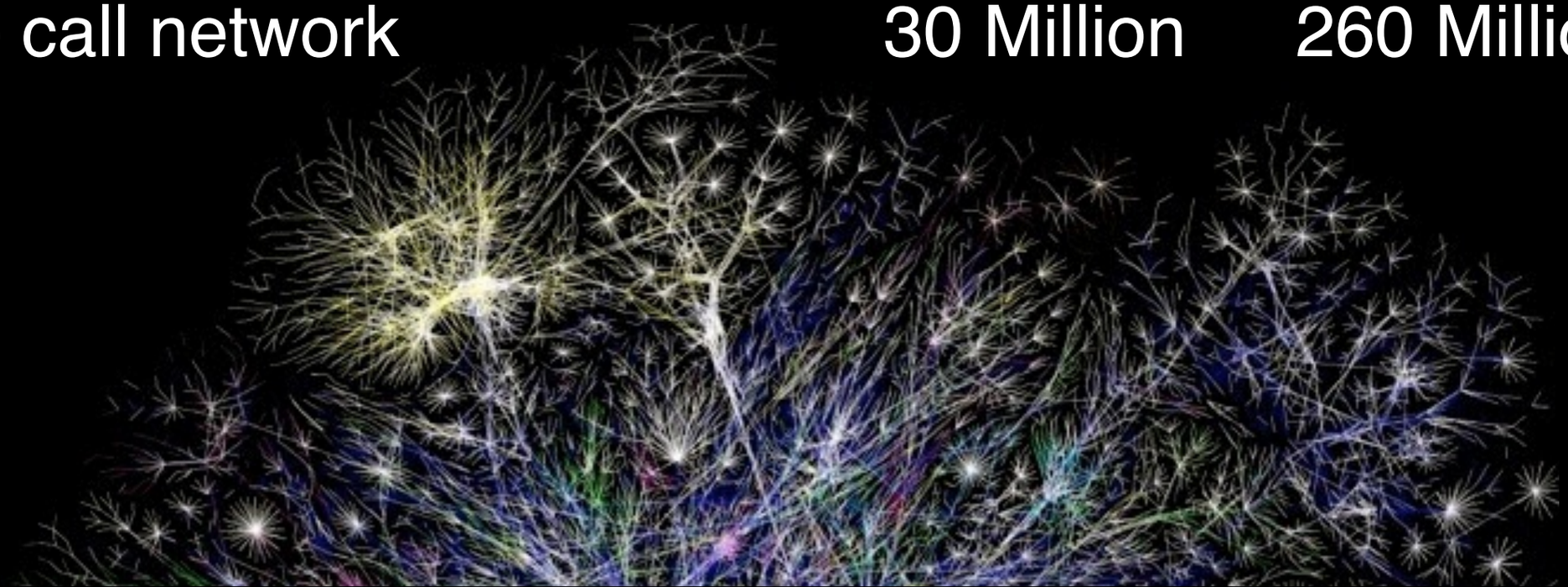
Protein-protein interactions

200 million possible interactions in human genome

Sources: www.selectscience.net www.phonedog.com www.mediabistro.com www.practicalecommerce.com/

Large Graphs I Analyzed

Graph	Nodes	Edges
YahooWeb	1.4 Billion	6 Billion
Symantec Machine-File Graph	1 Billion	**37 Billion**
Twitter	104 Million	3.7 Billion
Phone call network	30 Million	260 Million

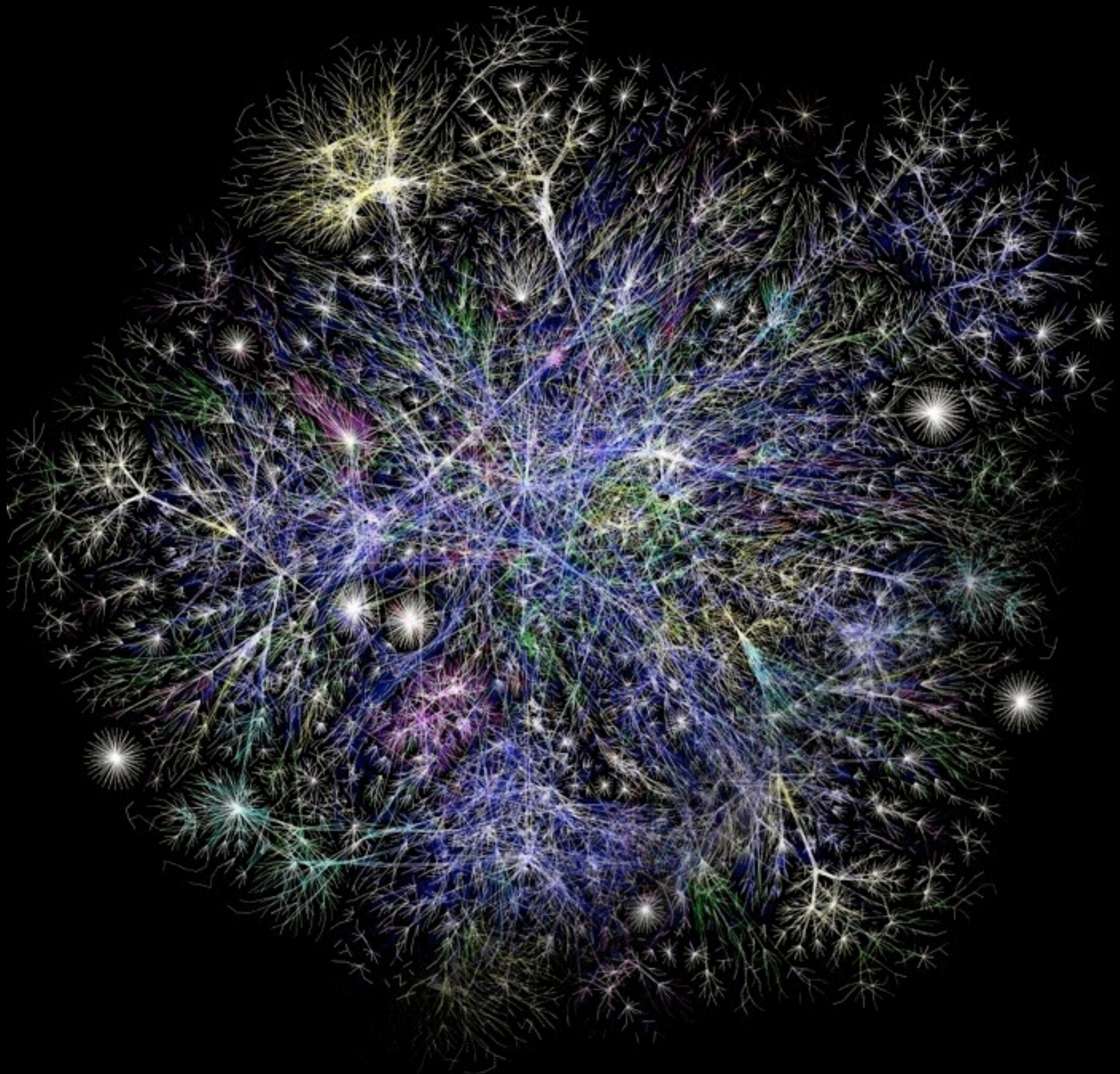


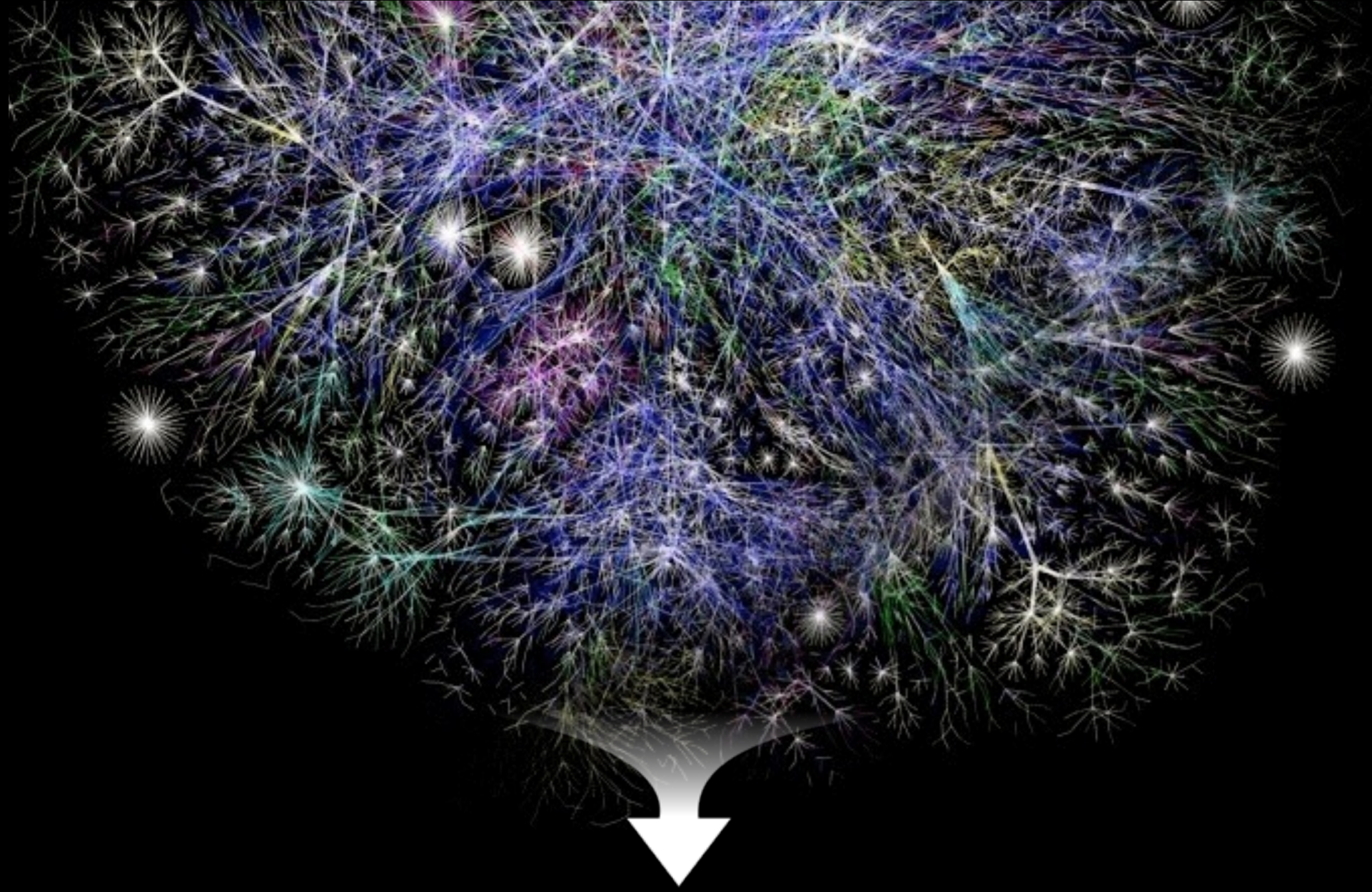
7

7±2

Number of items an average human holds in working memory

George Miller, 1956





7

Data

Insights

How to do that?

COMPUTATION + HUMAN INTUITION

How to do that?

COMPUTATION	INTERACTIVE VIS
Automatic	User-driven; iterative
Summarization, clustering, classification	Interaction, visualization
>Millions of nodes	Thousands of nodes

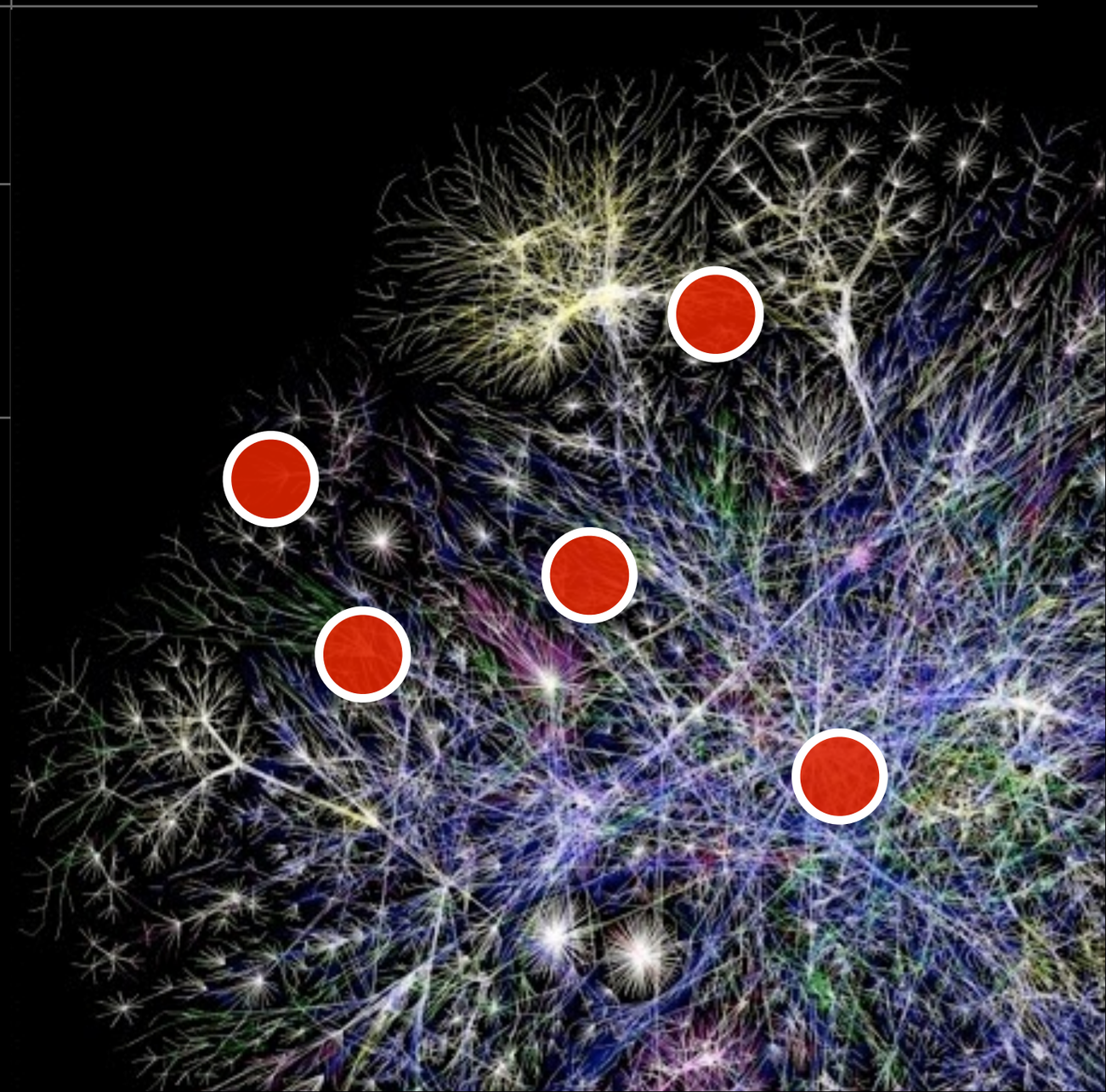
Both develop methods for making sense of network data

How to do that?

COMPUTATION	INTERACTIVE VIS
Automatic	
Summarization, clustering, classification	
>Millions of nodes	

How to do that?

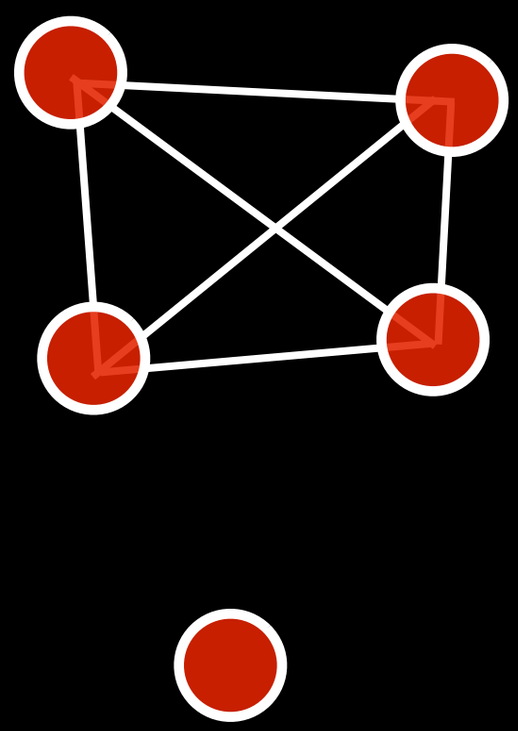
COMPUTATION	INTERACTIVE VIS
Automatic	
Summarization, clustering, classification	
>Millions of nodes	



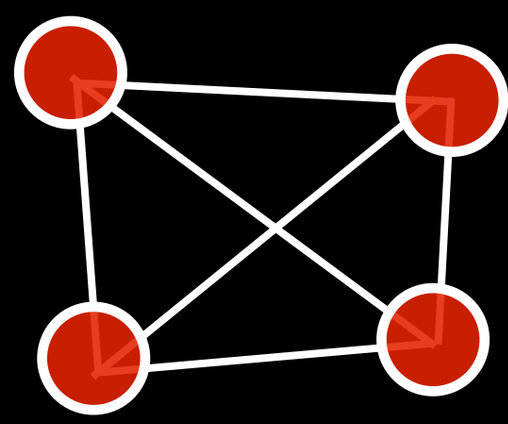
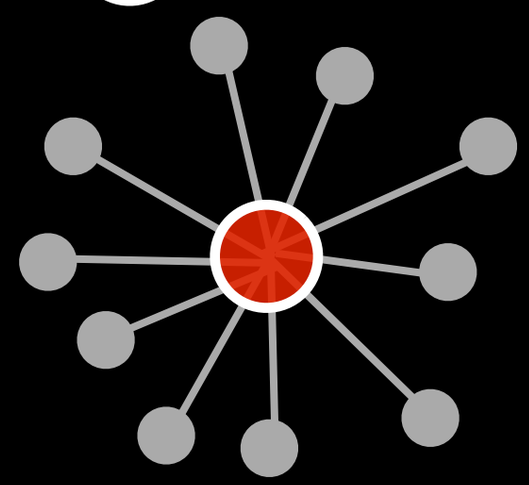
How to do that?

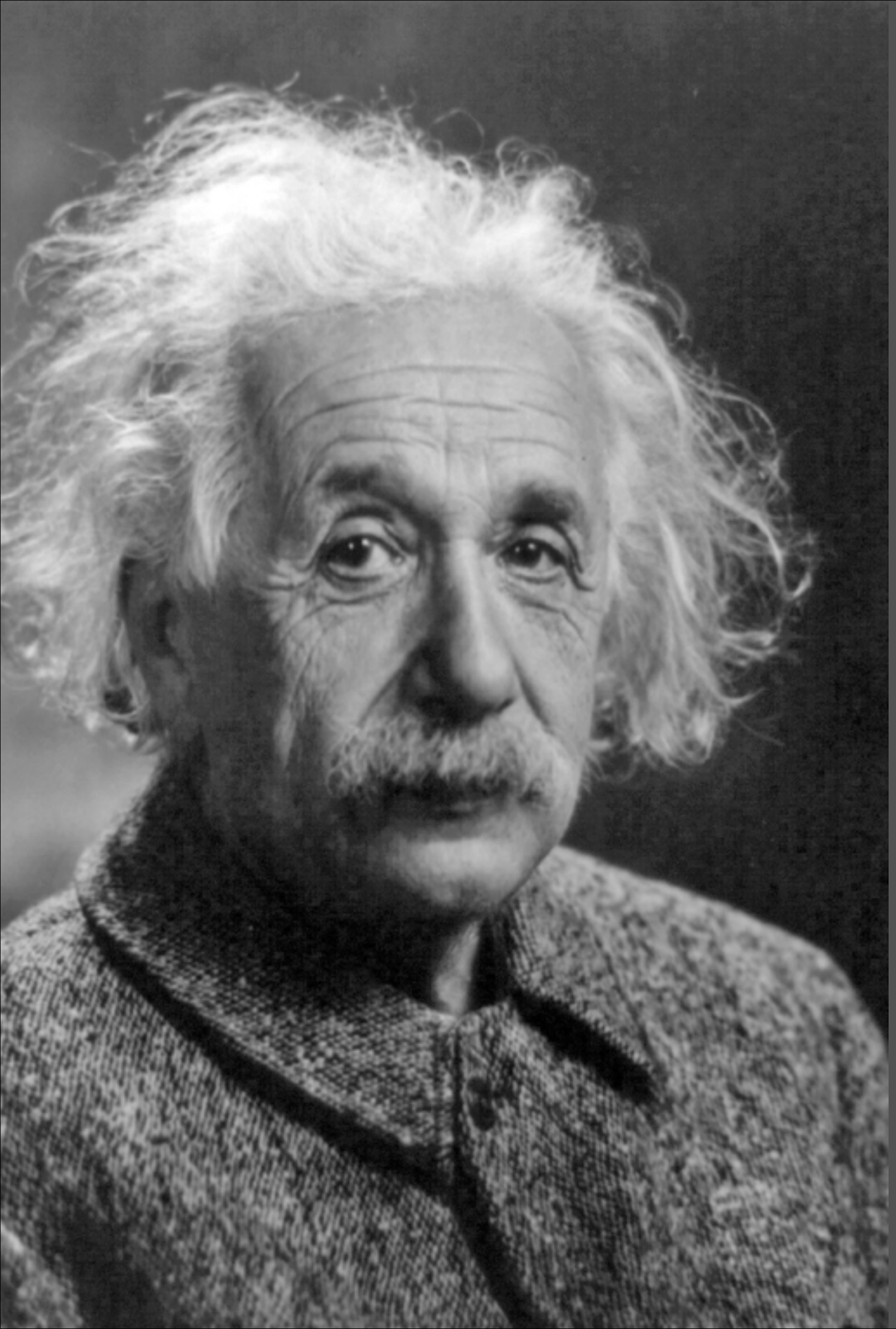
COMPUTATION	INTERACTIVE VIS
⋮	User-driven; iterative
	Interaction, visualization
	Thousands of nodes

How to do that?

COMPUTATION	INTERACTIVE VIS
	User-driven; iterative
	Interaction, visualization
	Thousands of nodes

How to do that?

COMPUTATION	INTERACTIVE VIS
	User-driven; iterative
	Interaction, visualization
	Thousands of nodes



"Computers are incredibly fast, accurate, and stupid.

Human beings are incredibly slow, inaccurate, and brilliant.

Together they are powerful beyond imagination."

“Essentially,

all models are wrong,

but some are useful”

George Box

Logistics

Course homepage	poloclub.gatech.edu/cse6242/
Discussion, Q&A, find teammates	Piazza (link on homepage, soon)
Submission	T-Square

Course Goals

- Learn **scalable** visual and computation techniques and tools, for typical data types
- Learn how to **combine** both kinds of methods (how they complement each other)
- Gain **practical** know-how
- Gain **breath** of knowledge

Course Expectation

- Overview of scalable visual and computation techniques and tools
- Gain knowledge & experience (useful for jobs, research)
- Experience with designing and developing an interactive analysis tool

Schedule

See course homepage

poloclub.gatech.edu/cse6242/

Grading

- 3-4 homework assignments (40%)
  - End-to-end analysis
  - Techniques (computation and vis)
  - Hadoop (+ other "big data" tools)
- Group project (50%) -- 3 to 4 people
- Participation (10%) -- in class, and on Piazza